Improving Mental Health Classifier Generalization with Pre-Diagnosis Data

Anonymous ACL submission

Abstract

Recent work has shown that classifiers for depression detection often fail to generalize to new datasets. Most NLP models for this task are built on datasets that use textual reports of a depression diagnosis (e.g., statements on social media) to identify diagnosed users; this approach allows for collection of large-scale datasets, but means that classifiers suffer from a self-report bias. Notably, models tend to capture features that typify direct discussion of mental health rather than more subtle indications of depression symptoms. In this paper, we explore the hypothesis that building classifiers using exclusively social media posts from before a user’s diagnosis will lead to less reliance on shortcuts and better generalization. We test our classifiers on a dataset that is based on an external survey rather than textual self-reports, and find that using pre-diagnosis data for training yields improved performance.

1 Introduction

In recent years, computational methods, including Natural Language Processing (NLP), have been applied to social media data, with the objective of learning about mental illness and improving mental healthcare (e.g., Coppersmith et al., 2015; Mitchell et al., 2015; Jamil et al., 2017; Cohen et al., 2020). A significant amount of work in this area focuses on the classification task of predicting mental health status from social media content. The main signals that have been used in order to infer mental health status for these classification tasks are listed in Table 1.

The practices of using self-reported diagnoses and community membership show promise from a machine learning perspective, in that data can be automatically labeled, and collecting datasets does not require participation from study “participants” in the form of surveys. This allows large datasets to be collected, which lend themselves well to deep learning methods. However, recent work has questioned the validity of these methods and their ability to generalize to new populations (Harrigian et al., 2020; Ernala et al., 2019).

Prior work has considered the fact that self-report bias is a significant obstacle in systems that rely on self-report (Harrigian et al., 2020; Chancellor et al., 2019). People who report their mental health diagnosis on social media are a subset of people on social media who have a mental health diagnosis, who are themselves a subset of people who have a mental health diagnosis (Figure 1). The behavior of people who are comfortable sharing their diagnosis on social media may differ in meaningful ways from those in the other two groups that are presented in Figure 1, which means that when we train and test classifiers only on self-report-based datasets, we are studying a subset of the population of interest. Furthermore, if our goals include early intervention and population-level monitoring, we should aim to do well at classifying all users who are symptomatic. The people included in datasets based on self-report differ from those who have a undiagnosed depression in that they have sought help and received a diagnosis from a professional. Their higher likelihood to be receiving treatment makes them differ from target populations in substantial ways, which may mean that features learned by classifiers will not generalize to all of those who are symptomatic. This mismatch
we extract the diagnosis timestamp for a subset of users. Finally, we build classifiers and test them on a dataset from another social media platform where mental health status is determined based on an external survey, rather than people’s behavior on the platform.\textsuperscript{1} We find that across a number of classifiers, generalization to the new population improves when using pre-diagnosis data for training.

\section{Related Work}

Mental health related textual data has been difficult to collect due to privacy issues and the cost of diagnosis. However, with the explosion of user-generated social media content, researchers have begun to consider how such content can be leveraged for the analysis of language usage related to mental health.

Prior work has mainly adopted two proxy signals to identify people with mental health conditions on social media platforms. The first and also the most popular proxy signal is a set of self-reported diagnosis patterns. Coppersmith et al. (2014) used patterns like “I was diagnosed with X” to identify more than 1,200 Twitter users with four mental health conditions (depression, bipolar disorder, PTSD, and SAD). Following this work, similar patterns were created for additional mental health conditions and different social media platforms (Coppersmith et al., 2015; Mitchell et al., 2015; Ernala et al., 2017; Yates et al., 2017; Cohan et al., 2018; Birnbaum et al., 2017). Based on these diagnosis patterns, some work further includes experts to verify the authenticity of identified diagnosed users (Mitchell et al., 2015; Ernala et al., 2017; Cohan et al., 2018; Birnbaum et al., 2017). The second proxy signal is the communities that users affiliate themselves with. McManus et al. (2015) identifies individuals with schizophrenia by checking if they follow the Twitter account @schizotribe. Jamil et al. (2017) identifies users who may have depression by searching within the #BellLetsTalk

\begin{table}[h]
\centering
\begin{tabular}{|l|l|l|}
\hline
Method & A person is labeled as depressed if... & Examples \\
\hline
Surveys and Healthcare Collaborations & their PHQ9 score on a survey reaches the level required to diagnose clinical depression. & Coughtry et al. (2013); Ernala et al. (2019) \\
Self-Reported Diagnosis & they post “I have been diagnosed with depression” (or similar) on social media. & Coppersmith et al. (2014); Cohan et al. (2018) \\
Community Membership & they join the r/depression Reddit community & Shen and Rudzicz (2017); Wolohan et al. (2018) \\
\hline
\end{tabular}
\caption{Signals used to infer mental health status for classification tasks. The lists of methods used for labeling and example papers are not exhaustive.}
\end{table}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure2.png}
\caption{By including all data from people who self-report, we introduce a mismatch between our training data and some real-life use cases.}
\end{figure}

\textsuperscript{1}Our code will be made publicly available.
We follow Cohan et al. (2018) by using self-reported diagnosis patterns to identify diagnosed users and collect corresponding control users based on their activity on Reddit. We look at all submissions and comments on Reddit from January 2006 to December 2019 using PushShift (Baumgartner et al., 2020). For convenience, we will use the term “post” to refer to both submissions and comments hereafter, and we do not distinguish them. Our dataset expands the Self-reported Mental Health Diagnoses (SMHD) dataset (Cohan et al., 2018) in three ways. First, we collect data from a longer time period. Second, we expand the list of mental health related keywords and subreddits used in SMHD. Third, our dataset includes self-report posts (discarded for training), which allow us to extract a diagnosis time (§4) and identify prediagnosis posts.

- **Diagnosed users** are identified by a list of self-reported diagnosis patterns from SMHD. An example of such a pattern is “I have been diagnosed with depression.” To reduce the false positive rate in retrieved data, another list of negative diagnosis patterns is used to remove users who do not have depression but are retrieved by the diagnosis patterns. Using such positive and negative patterns results in a set of diagnosed users with high precision (Cohan et al., 2018). Finally, we remove users who do not have at least 50 posts that are not about mental health (§3.1.2).

- **Control users** are identified for each diagnosed user based on their post activity. Specifically, we use three conditions for finding control users: (1) each control user must post in at least one common subreddit with the diagnosed user; (2) the number of posts of each paired control user and diagnosed user cannot deviate by a factor larger than two; and (3) control users cannot have any mental health related posts (§3.1.2). For each diagnosed user, we find nine corresponding control users.

### 3.1 Reddit Self-Report-Based Dataset (SELFREPORT)

#### 3.1.1 Data collection

We follow Cohan et al. (2018) by using self-reported diagnosis patterns to identify diagnosed users and collect corresponding control users based on mental health conditions. Participation in subreddits such as r/Acne and r/SuicideWatch are used as proxy signals to identify people with mental health conditions (Gkotsis et al., 2017; Shen and Rudzicz, 2017).

While identifying diagnosed people through proxy signals, some work analyzed the amount of time that passes between diagnosis and self-report (MacAvaney et al., 2018). However, they focus on classifying diagnosis recency and condition state, and do not study the impact of the time period spanned by user’s data on classifiers that predict mental health conditions.

Although social media platforms provide convenient access to a large amount of mental health data, previous work has identified several pitfalls when using such proxy signals to identify people with mental health conditions. Ernala et al. (2019) shows that people identified by proxy signals have different behaviors than people who are clinically diagnosed but do not post about mental health on social media. As a result, machine learning classifiers trained on such proxy signals cannot generalize to other populations that do not talk about mental health on social media. Harrigian et al. (2020) founds that models trained on data collected using a variety of proxy signals do not generalize across different social media platforms and proxies.

### 3.1.2 Mental health related data

We follow Cohan et al. (2018) by using a set of mental health terms and mental health subreddits to identify posts that relate to mental health. The posts including these terms are excluded when training classifiers to allow for better generalizability. In preliminary experiments, we found that the existing list does not include some terms and subreddits that closely relate to mental health (e.g., names of

---

2 Following Cohan et al. (2018) whose data collection process we built on, we will release code to collect data from Reddit. We cannot release the Twitter data due to the possibility of identifying individuals in the dataset (who may not have publicly shared their depression diagnosis) from their tweets.

3 e.g., “I’m not technically diagnosed.”
We collect 29,390 diagnosed users and 264,510 when training and testing) tend to be longer than Twitter. The dataset includes the Twitter handles ELF is measured in tokens.

gift card for completion of 4 surveys. 737 students completed the surveys, but we only include students who chose to provide active, informed consent form. 737 students completed the surveys, which underwent a full board review and was approved by the IRB at University of Anonymous. All participants in the study have signed an informed consent form. 737 students completed the surveys, but we only include students who chose to provide active, public Twitter handles. The students were given $50 worth of gift card for completion of 4 surveys.

antidepressants and r/2meirl4meirl4meirl tended to have high weights in linear classifiers, but using them for classification does not generalize to a population that does not explicitly talk about mental health. We therefore extend the existing list by adding a list of common antidepressants and additional mental health related subreddits.

### Data statistics

We collect 29,390 diagnosed users and 264,510 control users in total. We randomly split our dataset into train, validation, and test sets, which contain 70%, 15%, and 15% of the users, respectively. We present the statistics of the training set in Table 2. We observe that diagnosed users tend to have more and longer posts than control users. Furthermore, for the same set of diagnosed users, mental health related posts (which are excluded when training and testing) tend to be longer than other posts.

### Twitter Survey-Based Dataset (SURVEY)

In order to test our classifiers on a dataset that is not built on proxy signals, we use a dataset from Twitter. The dataset includes the Twitter handles from 210 students at a large US university.

<table>
<thead>
<tr>
<th># users</th>
<th># posts per user</th>
<th>post length</th>
<th># MH posts per user</th>
<th>MH post length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diagnosed users</td>
<td>20,573</td>
<td>753.8 (±1221.5)</td>
<td>40.0 (±76.6)</td>
<td>51.9 (±116.2)</td>
</tr>
<tr>
<td>Control users</td>
<td>185,157</td>
<td>497.3 (±957.2)</td>
<td>18.9 (±45.6)</td>
<td>—</td>
</tr>
</tbody>
</table>

Table 2: Statistics of SELFREPORT training set, which is based on self-reported depression diagnoses. Post length is measured in tokens.

<table>
<thead>
<tr>
<th># users</th>
<th># tweets per user</th>
<th>tweet length</th>
</tr>
</thead>
<tbody>
<tr>
<td>diagnosed-depression</td>
<td>32</td>
<td>1696.9 (±1481.7)</td>
</tr>
<tr>
<td>diagnosed-all</td>
<td>55</td>
<td>2974.8 (±9948.8)</td>
</tr>
<tr>
<td>control</td>
<td>138</td>
<td>1515.3 (±3913.0)</td>
</tr>
</tbody>
</table>

Table 3: Twitter SURVEY-based dataset statistics. Post length is measured in tokens.

are scraped using the Tweepy library. The students provided their Twitter handles in 2018 and 2019, and also completed a survey asking if the student had been diagnosed with a mental health condition. We split the students into three sets: students who state that they have diagnosed depression (diagnosed-depression), students who state that they have any mental health diagnosis (diagnosed-all), and students who state that they have never been diagnosed with any mental health conditions (control). Statistics of the dataset are displayed in Table 3. The dates of tweets in the dataset range from 2009 to 2020.

A limitation of our study is that this Twitter dataset is our only source of out-of-domain data that is not collected based on self-report. We would have preferred to test on multiple such datasets, but due to privacy concerns such data is very difficult to procure, and it is usually excluded entirely from NLP research on mental health.

### Diagnosis Timestamp Extraction

To study the temporal effect of diagnosis on depression classification methods, we extract the diagnosis timestamp for diagnosed users in the SELFREPORT dataset when possible. In the post in which users self-report their diagnosis, some users also share their diagnosis time. For such users, we extract their diagnosis timestamp with two-week precision or better from the text of their self-report post. To do this, we look only at the sentence where the self-reported diagnosis appears. From the sentence, we extract all time expressions that describe a DATE or TIME using SUTime (Chang and Manning, 2012). However, extracted time expressions do not necessarily describe the diagnosis time. We further get the dependency parsing tree of the sentence using spaCy (Honnibal and Montani, 2017), and we only include time expressions that can be reached from the self-reported diagnosis pattern in

---

4Community description: “For relatable posts that are too real for r/meirl or r/me_irl. Meaning jokes/posts about mental health issues and self depreciating humour.”


6The data was collected as part of a study that underwent a full board review and was approved by the IRB at University of Anonymous. All participants in the study have signed an informed consent form. 737 students completed the surveys, but we only include students who chose to provide active, public Twitter handles. The students were given $50 worth of gift card for completion of 4 surveys.

7https://www.tweepy.org/.

8Included are depression, anxiety, substance abuse disorder, personality disorder, eating disorder, attention disorder or learning disability.

9e.g., “I was diagnosed 3 months ago.”
If no time expression can be reached, we go to the parent of the diagnosis pattern and check again. Finally, we remove time expressions that are unlikely to be precise within a two-week period such as “several months ago” and “years ago.” Figure 3 shows an example where we can extract the precise timestamp “February.” An example for which we cannot extract a precise diagnosis time is “Anyway, in 2017, I had my depression diagnosis,” because “2017” is not precise on two-week level.

We are able to extract the diagnosis timestamp for 691 users (3.36% of total users) with two-week precision, and find that on average, there are 119.2 days between a user’s self-report and their diagnosis. To evaluate the accuracy of the extracted diagnosis times, we randomly sample 100 users from the training set and manually annotate their diagnosis time. Our method achieves 96.3% precision on sampled data while retaining 59.1% recall.

5 Experimental Setup and Results

To evaluate the effectiveness of using pre-diagnosis posts on improving models’ generalizability, we train classifiers using the SELFREPORT dataset from different time periods (§5.1) and test their performance across those periods in an in-domain setting. Then, we evaluate their ability to generalize to SURVEY and find that classifiers trained only on posts before diagnoses often outperform classifiers trained on all data in the transfer setting (§5.2).

5.1 Models

We consider four models ranging from Logistic Regression to Transformer-based language models (Vaswani et al., 2017; Ji et al., 2021). We search hyperparameters for each model on the validation set. Refer to Appendix A for details of the hyperparameter search and feature selection.

- **Logistic Regression**: we train two logistic regression models. The first one uses unigram and bigram TF-IDF features of the concatenated posts. The second one leverages Linguistic Inquiry and Word Count (LIWC) percentages (Pennebaker et al., 2015) of each post and uses their aggregation statistics (mean, variance, range, and quantile range) as features. In initial experiments, we used only the mean of LIWC values (as is common in NLP applications), but found that adding other aggregation statistics significantly improved the performance on both SELFREPORT and SURVEY.

- **FastText (Joulin et al., 2016)**: we train a FastText classification model using unigram and bigram features. We concatenate all of a user’s posts as the input.

- **MentalBERT (Ji et al., 2021)**: we utilize the contextual representations generated by a BERT-like (Devlin et al., 2019) language model that is adapted to mental health related content from Reddit (Ji et al., 2021). MentalBERT is trained on seven mental health related subreddits, so it does not overlap with SURVEY. Among the seven subreddits, two of them could contain posts that overlap with data in SELFREPORT. We feed the BERT representation of each post to a feed forward neural network and aggregate by max pooling to get the representation for each user.

We train each model on three subsets of data from SELFREPORT with different numbers of users and posts.
• **All-large** contains all posts from all users.
• **Pre-diagnosis** considers diagnosed users for whom we can extract the diagnosis timestamp and their corresponding control users. Users also need to have at least one post before their diagnoses. We only keep posts before diagnoses for diagnosed users and randomly sample the same percent of posts for control users.
• **All-small** contains the same set of users as **Pre-diagnosis**, but it includes all of their posts.

### 5.2 Results

#### 5.2.1 In domain performance

We first test classifiers on the **SELF REPORT** test set. We use the same definitions of **All-large**, **All-small**, and **Pre-diagnosis** on the test set. Table 4 shows the results. We can observe classifiers that are trained on all data and all users (**All-large**) achieve the best performance across all models on most of test cases including the pre-diagnosis cases, demonstrating their good fit for the population of users who self-report their depression diagnoses on Reddit. However, their performance drops significantly when testing on only pre-diagnosis posts, indicating they focus more on features that appear after diagnosis.

Table 5 illustrates the focus on features that are more likely to occur post-diagnosis. For example, **All-large** focuses on n-grams such as “meds” and “disorder” whereas **Pre-diagnosis** captures “insecure” and “my life”; similarly, **All-large** trained on LIWC features focuses on health- and anxiety-related words whereas **Pre-diagnosis** captures self-attentional focus, as indicated by first person singular pronouns (self preoccupation is known to relate with people’s psychological status (Pennebaker, 2004)). Although these features are indicative of a depression diagnosis for this specific population, as we show in §5.2.2, these post-diagnosis features fail to generalize to the broader population. Methods such as filtering lists of words that are directly related to mental health (as we do in §3.1) help to reduce reliance on these n-grams, but crafting these lists requires significant manual effort and subjective decisions. Changing the time period of data used reduces subjectivity and helps to filter out these features.

Next, we explore how the performance changes when we vary the time period covered by the test data. Concretely, we evaluate classifiers on diagnosed users using their posts before a specific time point. To exclude the influence of other factors, we test on the same set of diagnosed users who have at least one post 90 days before their diagnoses, and we down-sample users’ posts to keep the number of posts unchanged for different test time points. We also include the corresponding control users in the test set and down-sample their posts proportionally to the number of posts of diagnosed users. As shown in Figure 4, **Pre-diagnosis** has a relatively consistent performance for different time ranges.
Figure 4: F1 score for diagnosed users tested on SELFREPORT (average of three runs). Shaded area shows the 95% confidence interval. For diagnosed users, we only consider posts before a certain time point (x axis), and we down-sample the posts so that each user has the same number of posts at each test point. All results are on the same set of users who have at least one post 90 days before their diagnoses.

Table 6: Slope of the fitted lines in Figure 4 (on the order of 10\(^{-5}\)). The largest value for each model is in bold, and reflects the focus on post-diagnosis features.

<table>
<thead>
<tr>
<th></th>
<th>TF-IDF</th>
<th>LIWC</th>
<th>FastText</th>
<th>MentalBERT</th>
</tr>
</thead>
<tbody>
<tr>
<td>All-large</td>
<td>12.83</td>
<td>20.04</td>
<td>19.71</td>
<td>34.77</td>
</tr>
<tr>
<td>All-small</td>
<td>3.48</td>
<td>19.45</td>
<td>15.04</td>
<td>14.85</td>
</tr>
<tr>
<td>Pre-diagnosis</td>
<td>2.59</td>
<td>12.44</td>
<td>4.32</td>
<td>25.02</td>
</tr>
</tbody>
</table>

5.2.2 Out-of-domain performance

Next, we evaluate classifiers on SURVEY. SURVEY represents a broader population in that it does not use self-reported diagnoses to identify diagnosed users; this means that users are less likely to explicitly mention their mental health. We only use it for testing purposes, and we include all tweets for each user. The results are presented in Table 7. We first notice that Pre-diagnosis classifiers achieve the best performance for five out of eight models; notably, when training on the same set and number of users (All-small vs. Pre-diagnosis), there is an improvement for seven of the eight models. This verifies our hypothesis that training on pre-diagnosis posts generalizes better to the broader population. We noted that training on All-small sometimes unexpectedly improves upon All-large; it is possible that there are unobserved demographic overlaps between All-small and SURVEY (a relatively homogenous group at one university) that cause this to occur. Given that, we focus primarily on the direct comparison between All-small and Pre-diagnosis.

We also observe that for a broader set of mental health conditions (diagnosed-all, which includes users who have diagnosed anxiety and eating disorders in addition to users with diagnosed depression), training on pre-diagnosis posts provides stronger generalizability, as shown by larger gaps between Pre-diagnosis classifiers and All-small classifiers. We believe this generalization comes from some common symptoms shared by those with anxiety and depression (Hanson, 2019). We were somewhat surprised to see a pattern of higher scores on the diagnosed-all dataset than the diagnosed-depression dataset, given that the users in the diagnosed-depression dataset all share a diagnosis with the users in the SELFREPORT dataset. From Table 3, we note that users in the diagnosed-all dataset tend to have more tweets; this could increase the likelihood that at least some of their tweets contain signals that are indicative of their symptoms.

Finally, we acknowledge that some of our best results come from using MentalBERT and FastText models with all data (All-large). With immense computational resources, the MentalBERT model is presumably able to extract generaliz-
Table 7: F1 score for diagnosed users on SURVEY (mean of three runs; error shows standard deviation). \( \Delta \) means Pre-diagnosis – All-small. Diagnosed-depression contains diagnosed-depression and control users. Diagnosed-all contains diagnosed-all and control users. Best results are in bold. Random baseline achieves 27.35 for Diagnosed-depression and 36.30 for Diagnosed-all.

<table>
<thead>
<tr>
<th></th>
<th>TF-IDF</th>
<th>LIWC</th>
<th>FastText</th>
<th>MentalBERT</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Diagnosed-depression</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Random</td>
<td>27.35</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All-large</td>
<td>41.18 ± 0.00</td>
<td>40.86 ± 0.00</td>
<td>38.41 ± 2.43</td>
<td>42.02 ± 6.44</td>
</tr>
<tr>
<td>All-small</td>
<td>42.86 ± 0.23</td>
<td>41.02 ± 0.32</td>
<td>35.63 ± 3.04</td>
<td>42.88 ± 2.83</td>
</tr>
<tr>
<td>Pre-diagnosis</td>
<td>48.11 ± 0.20</td>
<td>49.02 ± 0.36</td>
<td>41.69 ± 0.67</td>
<td>43.96 ± 2.31</td>
</tr>
<tr>
<td><strong>Diagnosed-all</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Random</td>
<td>36.30</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All-large</td>
<td>40.33 ± 2.31</td>
<td>40.04 ± 2.57</td>
<td>38.75 ± 0.78</td>
<td>38.07 ± 2.30</td>
</tr>
<tr>
<td>All-small</td>
<td>41.64 ± 0.00</td>
<td>40.04 ± 0.00</td>
<td>38.41 ± 2.43</td>
<td>42.02 ± 6.44</td>
</tr>
<tr>
<td>Pre-diagnosis</td>
<td>48.11 ± 0.20</td>
<td>49.02 ± 0.36</td>
<td>41.69 ± 0.67</td>
<td>43.96 ± 2.31</td>
</tr>
</tbody>
</table>

6 Discussion

In our experiments, we find that classifiers using exclusively Pre-diagnosis data tend to outperform classifiers using all data from all users with self-reported depression when tested on a population that does not self-report their diagnosis online. While we did not see the same trend on in-domain data, we did find that when posts are downsampled such that classifiers are tested using the same number of posts per user, the Pre-diagnosis model achieves comparable results to the others that have been trained on far more data.

This is notable in that (a) it shows that a more focused approach in the data curation stage can help with the generalization issues noted by Harrigan et al. (2020) and Ernala et al. (2019), and (b) we can achieve strong results with far less data, reducing the necessary computational resources (for the MentalBERT model, Pre-diagnosis takes less than 1% of the training time of the All-large model on the same device). Our study shows that relying exclusively on big data is not enough to build effective classifiers; rather, data quality is of the utmost importance. Improving data quality is possible by re-examining assumptions in the data-curation process.

To the best of our knowledge, this is the first study to investigate how temporal factors affect mental health classifiers. It opens up the possibility of exploring the same phenomena on other mental health diagnoses (e.g. anxiety). Additionally, it reinforces the need to consider temporal variation in any classification problem that aims to classify people based on text they write over time but considers their behavior to be static.

By focusing on pre-diagnosis data, we also demonstrate an approach that would likely generalize particularly well to people who are not yet diagnosed and need to be connected to help (Figure 2). While the SURVEY dataset contains no indication of when each user was diagnosed, it is likely that the improvements observed when using pre-diagnosis data would be enlarged if we only had that data at test time. In the future, we believe that precise data-curation methods can be used in conjunction with modeling techniques that aim to improve generalizability (Lee et al., 2021) to build more robust classifiers.

7 Conclusion

In this paper, we proposed and validated the hypothesis that using only data from before users are diagnosed can improve the generalization of mental health classifiers. In particular, we focus on addressing self-report bias; in datasets based on social media, diagnoses are often inferred from what users say about themselves, meaning that classifiers built on their data focus on a specific subset of people (Figure 1). When validating our hypothesis, we tested our classifiers on a dataset that is collected in a way that avoids this bias. Our results showed that reconsidering the set of data that we use for training can lead to improved performance, especially for models that focus on social phenomena.
8 Ethical Considerations

While all of the data we use for training is publicly available, we acknowledge that people’s mental health conditions are highly personal. While the users in our training set all shared their depression diagnosis on an online forum, we acknowledge that special care should be taken with such data considering the sensitivity of the subject. As has been done in prior studies, we removed identifiers such as Reddit usernames from our personal copy of the data, and make no attempt to ascertain any information about the users who comprise our dataset beyond what is written in their Reddit posts. The study that resulted in the Twitter dataset received full IRB approval from our institution; personal identifiers such as Twitter usernames were scrubbed from the dataset. Ethics around health-related social media data are explored in more detail in Benton et al. (2017).

In our work, we show one method that improves generalizability of depression diagnosis classifiers (with simpler models) to a population of people who may not explicitly discuss mental health. While our method improves upon the baseline, the results do not suggest that such a model is ready for real-world deployment. The task of detecting depression from text is very challenging, and our results on out-of-domain data that is collected without using self-reported diagnoses show that we still have a long way to go with respect to accuracy on populations that differ from those that we see in our training data (regardless of how that data is sampled).

However, the more important question to ask may be how such a classification system should be used if accuracy reaches an acceptable threshold and how to define that threshold for various potential applications. A very accurate depression classification system could be used for good: monitoring population-level depression (e.g., Wolohan (2020)), routing counselors to those with the most need in resource-constrained settings (e.g., as recommended by Bantilan et al. (2020)), opt-in prompts to receive counseling on college campuses if classifiers see symptoms developing, or opt-in monitoring for people who are already receiving counseling. However, the same systems could also be used for nefarious purposes, such as denying jobs to people whose mental health status is inferred from their social media posts. This would be illegal in the United States, but it may not be in all countries, and an action being illegal does not eliminate the risk of it occurring. While out-of-scope for this paper, the question of how mental health classifiers should be used and which classification setups will most benefit society while reducing harm should be considered more thoroughly by the community, with active involvement from mental health practitioners. To the best of our knowledge, these considerations have been understudied in the NLP community; the few exceptions that focus on the ethical tensions surrounding mental health classifiers have appeared outside of NLP (Chancellor et al., 2019). We hope that the community will consider and participate in inner-disciplinary work that directly considers how mental health classification models can be deployed; one recent example of such work is Cohen et al. (2020).
References


Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media.


Jolene Hanson. 2019. Identifying anxiety, depression signs.


Table 8: Hyperparameters used to train TF-IDF and LIWC models. max_df and min_df mean the maximum and minimum document frequency. $|D|$ is the number of documents in the training set.

<table>
<thead>
<tr>
<th>Hyperparameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>regularization strength</td>
<td>searched [0.3, 1, 3]</td>
</tr>
<tr>
<td>TF-IDF min_df</td>
<td>5</td>
</tr>
<tr>
<td>TF-IDF max_df</td>
<td>$0.7 \times</td>
</tr>
<tr>
<td>TF-IDF max # features</td>
<td>100000</td>
</tr>
</tbody>
</table>

Table 9: Hyperparameters used to train FastText models.

<table>
<thead>
<tr>
<th>Hyperparameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>minimum word occurrence</td>
<td>10</td>
</tr>
<tr>
<td>features</td>
<td>unigram and bigram</td>
</tr>
<tr>
<td>learning rate searched</td>
<td>[0.4, 0.5, 0.6, 0.7]</td>
</tr>
<tr>
<td># epochs searched</td>
<td>[40, 50, 60, 70]</td>
</tr>
</tbody>
</table>

Table 10: Hyperparameters used to train MentalBERT models.

<table>
<thead>
<tr>
<th>Hyperparameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>number of epochs</td>
<td>20</td>
</tr>
<tr>
<td>patience</td>
<td>3</td>
</tr>
<tr>
<td>maximum learning rate</td>
<td>searched [0.0001, 0.001]</td>
</tr>
<tr>
<td>learning rate scheduler</td>
<td>linear decay</td>
</tr>
<tr>
<td>optimizer</td>
<td>Adam (Kingma and Ba, 2017)</td>
</tr>
<tr>
<td>weight decay searched</td>
<td>[0.0005, 0.005]</td>
</tr>
<tr>
<td>Adam beta weights</td>
<td>0.9, 0.999</td>
</tr>
<tr>
<td># FNN layer</td>
<td>3</td>
</tr>
<tr>
<td>dimension of FNN</td>
<td>[512, 128, 128]</td>
</tr>
<tr>
<td>dropout in FNN</td>
<td>0.1</td>
</tr>
</tbody>
</table>

A.3 MentalBERT Models

We use the MentalBERT model to generate representations for each user post (Ji et al., 2021). The model is trained on user posts from seven mental-health related subreddits: r/depression, r/SuicideWatch, r/Anxiety, r/offmychest, r/bipolar, r/mentalillness/, and r/mentalhealth. Among these seven subreddits, r/offmychest and r/mentalillness/ might contain posts that overlap with data in SELF-REPORT. For efficiency, we use the fixed representations ([CLS] token) generated by MentalBERT and do not fine-tune it. We pass the MentalBERT representations to a Feedforward Neural Network (FNN) and use max pooling to get aggregated user representation.

We train all models on a GeForce RTX 2080 Ti GPU. Generating the fixed MentalBERT representations takes around 29 hours. The training time is 8 hours for the All-large model and 3 minutes for the All-small and Pre-diagnosis models. The hyperparameters for the MentalBERT classifiers are in Table 10.

---

13https://fasttext.cc